Volume Title
ASP Conference Series, Vol. **Volume Number**
Author

Automated curation of infra-red imaging data in the WFCAM and VISTA Science Archives

Nicholas Cross[1], Ross Collins[1], Eckhard Sutorius[1], Nigel Hambly[1], Rob Blake[1], and Mike Read[1].

[1]*Scottish Universities' Physics Alliance (SUPA),Institute for Astronomy, School of Physics, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK*

Abstract. The two fastest near infrared survey telescopes are UKIRT-WFCAM and VISTA. The data from both these instruments are being archived by Wide Field Astronomy Unit (WFAU) at the IfA, Edinburgh, using the same curation pipeline, with some instrument specific processing. The final catalogues from these surveys will contain many tens of billions of detections.

Data are taken for a range of large surveys and smaller PI programmes. The surveys vary from shallow hemisphere surveys to ultra deep single pointings with hundreds of individual epochs, each with a wide range of scientific goals, leading to a wide range of products and database tables being created. Processing of the main surveys must allow for the inclusion of specific high-level requirements from the survey teams, but automation reduces the amount of work by archive operators allowing a higher curation efficiency. The decision making processes which drive the curation pipeline are a crucial element for efficient archiving. This paper describes the main issues involved in automating the pipeline.

1. The WFCAM and VISTA Science Archives

The WFCAM and VISTA Science Archives (Hambly et al. 2008) are the main access for data from WFCAM (Casali et al. 2007) and VISTA (Emerson & Sutherland 2010). The majority of time on both instruments is spent on large surveys: UKIDSS (Lawrence et al. 2007), and the VISTA Public Surveys (Arnaboldi et al. 2007). There are also a range of smaller Principal Investigator (PI) programmes allocated by the Telescope Allocation Committees each semester that require curating. We run the same set of tasks on all the surveys and programmes, with the amount of processing in each task dependent on the type of programme. For instance a wide survey will spend more time on band-merging and neighbour tables, but a deep survey will spend more time on deep stack creation and multi-epoch tables. PI programmes are set up completely automatically[1] because of the large number of programmes, but surveys are set up in a semi-automatic way receiving special instructions, quality control and sometimes additional products from the science teams. PI programmes usually obtain all their data in one observing semester and are processed completely at the end of the semester, so new releases will be due to software or calibration improvements necessitating a complete reprocessing. Surveys build up data over many semesters and will be appended to, as

[1]Occasionally we have manually grouped together several related PI programmes from different semesters before automatic processing.

well as occasionally reprocessed. These different scenarios have to be factored into the pipeline control.

2. Overview of data pipeline

Data consisting of images and catalogues are first transferred to WFAU by the Cambridge Astronomy Survey Unit (CASU) who process each observing block and calibrate the data. We ingest these into the science archives along with any external data or quality control provided by the science teams for public surveys. The automated curation pipeline is then run, executing the following tasks:

- Quality Control
- Programme setup (using **ProgrammeBuilder** class)
- Creation and ingestion of deep products and catalogues
- Creation of band-merged Source table from deepest products
- Recalibration of each epoch
- Creation of band-merged catalogues for each epoch
- Creation of neighbour tables
- Creation of synoptic tables for light-curves and variability analysis

The dataflow plan for automated curation of a single programme is shown in Fig 1a. The automated pipeline has changed in a couple of ways since Collins et al. (2009). We have removed the distinction between deep and shallow programmes in a way that also allows us to do an easier comparison between the curation and data tables and removes the need to copy data. Appending either in width (more pointings) or depth (more epochs) is much easier. We have also designed a much more sophisticated SQL schema template that can be used by both surveys and PI programmes so that only one template is needed per instrument. These changes have significantly increased the amount of automation of public surveys so that many of the curation tables are filled in the same way as PI programmes.

3. Setting up a programme

The curation tables that control the pipeline are filled by the **ProgrammeBuilder** class, see Fig 1b. The image metadata is used to group pawprint frames by position, filter, microstepping and in the case of VISTA position angle and offset position within a standard tile. Unique pointings are found by grouping by position alone and then products are found for each filter in these pointings. The information for each pawprint product is put into the `RequiredStack` table. For VISTA, the pawprint stacks are grouped into tiles in `RequiredTile` and the two tables are linked via `ProductLinks`. For each filter used, the number of epochs at each pointing is found, and this is used to determine whether multi-epoch tables are required. This information goes into `RequiredFilters`, `Programme` and `RequiredNeighbours`.

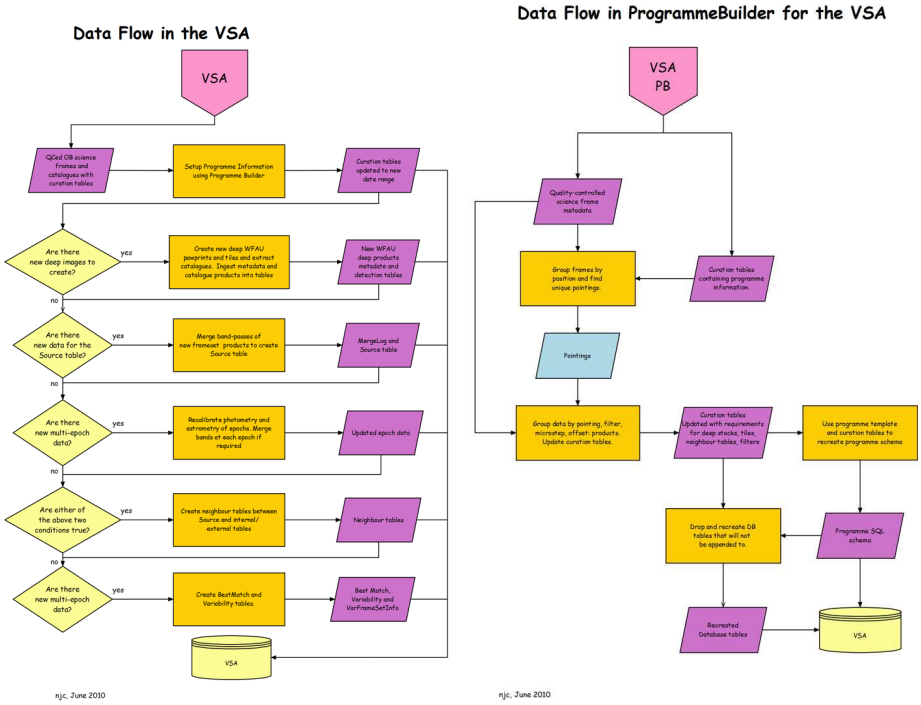

Figure 1. Left (a): Overall dataflow for WSA or VSA automated processing. Right (b): Dataflow for the ProgrammeBuilder class. In both figures, the rectangular yellow boxes represent tasks, the parallelograms represent data (light blue for temporary and purple for permanent) and the light yellow diamonds represent control structures. The cylinder represents the database that the permanent data are ingested into and the pentagon represents the pipeline that the dataflow is embedded in. In Fig a), processing of each task to create a permanent data object is determined by a control structure. New data feeds into the control for the next task. Each control structure compares expected products, based on the sorting and grouping of data in Fig b), to actual created products and before determining what still needs to be created.

`RequiredNeighbours` contains the list of all the neighbour and cross-match tables which need to be created. In the case of PI programmes, a common set of internal neighbour tables and cross-matches to all-sky surveys is produced and in addition surveys are cross-matched with wide range of specified external surveys.

Once these curation tables are setup, a programme specific schema is created using them and a template SQL schema. The template schema is composed of SQL definitions for different attributes, substitution strings and control structures. We give an example below from the template for VISTA `Source` tables:

```
++c:a
**s*&a&m&b&Pnt      real not null,      --/D Point source colour
&As&-&Bs& (using aperMag3)  --/U mag  --/C PHOT_COLOR  --/Q
&a&AperMag3,&b&AperMag3  --/N -0.9999995e9  --/G
allSource::colours
==c:a
```

The + + c : a line is a control structure which repeats each subsequent line for all the programme filters until the == c : a line for all colour combinations in the survey. The ∗ ∗ s∗ structure controls which table(s) each line goes into when several narrow tables are created for curation purposes and subsequently joined at release into one table. Lines will only go into the `Source` table, but not the `MergeSource`. &a&, &A&, &b& and &B& are substitution strings, where a and b refer to the first and second filter in a colour respectively and a and A refer to lower and upper case respectively. When the template is processed for the VISTA-VMC (Cioni, M.-R. et al. 2011, in prep.),which contains Y, J and Ks band data, the following piece of schema is produced:

```
ymjPnt      real not null,      --/D Point source colour Y-J
(using aperMag3)  --/U mag  --/C PHOT_COLOR  --/Q yAperMag3,
jAperMag3  --/N -0.9999995e9  --/G allSource::colours
jmksPnt      real not null,      --/D Point source colour J-Ks
(using aperMag3) --/U mag  --/C PHOT_COLOR  --/Q jAperMag3,
ksAperMag3  --/N -0.9999995e9  --/G allSource::colours
```

The SQL schema is used to create the database with the correct tables, control the code that produces the table data and create a schema browser and glossary for scientists. Having a single template and control structures reduces the need to repeat the SQL, which makes it much easier to update and maintain.

4. Triggering the automated pipeline

Each task is triggered by comparing the curation tables with the data tables. For instance, the deep product curation task will compare the products specified in `RequiredStack` and `RequiredTile` with data in the tables `Multiframe` and `ProgrammeFrame` matching on the programme, the product identifier and the release number. If all required products have been created then the pipeline moves onto the next task. If not, the remaining products will be created. Thus the pipeline can be restarted easily if there is a network error or software bug. Other curation tasks are similarly triggered and the data tables are updated at the end of each task. Log files are produced and curation history tables are kept up to date, so any failures can easily be identified.

References

Arnaboldi, M., Neeser, M. J., Parker, L. C., & et al. 2007, The Messenger, 127, 28

Casali, M., Adamson, A., Alves de Oliveira, C., & et al. 2007, A&A, 467, 777

Collins, R., Cross, N. J., Sutorius, E., & et al. 2009, in Astronomical Society of the Pacific Conference Series, edited by D. A. Bohlender, D. Durand, & P. Dowler, vol. 411 of Astronomical Society of the Pacific Conference Series, 226

Emerson, J. P., & Sutherland, W. J. 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, vol. 7733 of Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference

Hambly, N. C., Collins, R. S., Cross, N. J. G., & et al. 2008, MNRAS, 384, 637. 0711.3593

Lawrence, A., Warren, S. J., Almaini, O., & et al. 2007, MNRAS, 379, 1599. arXiv:astro-ph/0604426